SAINT: **S**elf-**A**ttention Augmented **I**nception-Inside-Inception **Net**work Improves Protein Secondary Structure Prediction

Mostofa Rafid Uddin[1,†], Sazan Mahbub[1,†], M Saifur Rahman[1], and Md Shamsuzzoha Bayzid[1,*]

[1]Department of Computer Science and Engineering
Bangladesh University of Engineering and Technology
Dhaka-1205, Bangladesh
[†]These authors contributed equally to this work
[*]Corresponding author: shams_bayzid@cse.buet.ac.bd

Abstract

Motivation: Protein structures provide basic insight into how they can interact with other proteins, their functions and biological roles in an organism. Experimental methods (e.g., X-ray crystallography, nuclear magnetic resonance spectroscopy) for predicting the secondary structure (SS) of proteins are very expensive and time consuming. Therefore, developing efficient computational approaches for predicting the secondary structure of protein is of utmost importance. Advances in developing highly accurate SS prediction methods have mostly been focused on 3-class (Q3) structure prediction. However, 8-class (Q8) resolution of secondary structure contains more useful information and is much more challenging than the Q3 prediction.

Results: We present SAINT, a highly accurate method for Q8 structure prediction, which incorporates self-attention mechanism (a concept from natural language processing) with the Deep Inception-Inside-Inception (Deep3I) network in order to effectively capture both the *short-range* and *long-range interactions* among the amino acid residues. SAINT offers a more interpretable framework than the typical black-box deep neural network methods. Through an extensive evaluation study, we report the performance of SAINT in comparison with the existing best methods on a collection of benchmark datasets, namely, TEST2016, TEST2018, CASP12 and CASP13. Our results suggest that self-attention mechanism improves the prediction accuracy and outperforms the existing best alternate methods. SAINT is the first of its kind and offers the best known Q8 accuracy. Thus, we believe SAINT represents a major step towards the accurate and reliable prediction of secondary structures of proteins.

Availability: SAINT is freely available as an open source project at `https://github.com/SAINTProtein/SAINT`.

Keywords: Protein secondary structure, deep learning, self-attention.

1 Introduction

Proteins are bio-molecules made of long chains of amino acid residues connected by peptide bonds. The functions of proteins are usually determined by the their tertiary structure and for determining the tertiary structure and related properties, the secondary structure information is crucial. Protein structure can be experimentally determined by X-ray crystallography and multi-dimensional magnetic resonance in laboratory, but these methods are very costly and time consuming and are yet to be consistent with the proliferation of protein sequence data [1]. Thus, the proteins with known primary sequence continue to outnumber the proteins with experimentally determined secondary structures. The structural properties of a protein depend on its primary sequence [2–5], yet it remains as a difficult task to accurately determine the secondary and tertiary structures of proteins. Hence, the problem of predicting the structures of a protein – given its primary sequence – is crucially important and remains as one of the greatest challenges in computational biology.

Secondary structure – a conformation of the local structure of the polypeptide backbone – prediction dates back to the work of Pauling and Corey in 1951 [6]. The secondary structures of proteins are traditionally characterized as 3 states (Q3): helix (H), strand (E), and coil (C). Afterwards, a more fine-grained characterization of the secondary structures was proposed [7] for more precise information by extending the three states into eight states (Q8): α-helix (H), 3_{10}-helix (G), π-helix (I), β-strand (E), isolated β-bridge (B), turn (T), bend (S), and Others (C). Q8 prediction is more challenging and can reveal more precise and high resolution on the structural properties of proteins.

Protein secondary structure prediction is an extensively studied field of research [8–30]. Developing computational approaches (especially using machine learning techniques) for 3-state SS prediction has a long history which dates back to the works of Qian & Sejnowski [8] and Holley & Karplus [9] who first used neural networks to predict SS. In the 1980s, only statistical model based methods were used on raw sequence data which could ensure Q3 accuracy merely below 60%. Afterwards, significant improvement was achieved [10–12] by leveraging the evolutionary information such as the position-specific score matrices (PSSM) derived from multiple sequence alignments. Subsequently, many machine learning methods have been developed for Q3 prediction which include support vector machines (SVM) [13–15, 31], probabilistic graphical models [16, 32, 33], hidden Markov models [17, 18], bidirectional recurrent neural networks [19–22, 34, 35], and other deep learning frameworks [23, 36].

The performance of Q3 prediction methods has approached the postulated theoretical limit [24]. At the same time, there has now been a growing awareness that 8-state prediction can reveal more valuable structural properties. Accurate 8-state secondary structures predictions can reduce the search space in template-free protein tertiary structure modeling by restricting the variations of backbone dihedral angles within a small range according to the Ramachandran plots [37, 38]. Also, differentiation among 3_{10} helix, α-helix, and π-helix in secondary structure prediction helps to assign residues and fit protein structure models in cryo-electron microscopy density maps [38, 39]. As such, the interest of the research community has recently shifted from Q3 prediction to relatively more challenging Q8 prediction. Quite a few deep learning methods for Q8 prediction have been proposed over the last few years [19, 25, 26, 28–30, 40]. To the best of our knowledge,

the first notable success in Q8 prediction methods was SSpro8 [19] which was published in 2002 and achieved 63.5% Q8 accuracy on the benchmark CB513 dataset [41], 64.9% on CASP10 and 65.6% on CASP11 [25]. Later in 2011, RaptorX-SS8 [40], another 8 state predictor using conditional neural fields, surpassed SSpro8 by demonstrating 64.9% Q8 accuracy on CB513. In 2014, Zhou and Troyanskaya [26] highlighted the challenges in 8-state prediction and obtained 66.4% Q8 accuracy on CB513 dataset using deep generative stochastic network (GSN). Some of the notable subsequent works include deep conditional random fields (DeepCNF) [25], cascaded convolutional and recurrent neural network (DCRNN) [27], next-step conditioned deep convolutional neural network (NC-CNN) [28], multi-scale CNN with highway (CNNH_PSS) [29], DeepACLSTM [42] with an asymmetric convolutional neural networks (ACNNs) combined with bidirectional long short-term memory (BLSTM), deep inception-inside-inception (Deep3I) network named MUFOLD-SS [30], CNN and Bidirectional LSTM based network NetSurfP-2.0 [43], and SPOT-1D [44] which is an ensemble of of hybrid models consisting of Residual Convolutional Neural Networks (ResNet) and 2-Dimensional Bidirectional Residual LSTM Networks (2D-BRLSTM). While most of the methods use sequence data and sequence profiles obtained from Position Specific Scoring Matrix (PSSM) as features, the more recent methods, such as, MUFOLD-SS, CRRNN, NetSurfP-2.0, and SPOT-1D leveraged HMM profiles and physicochemical properties of residues as well. The most recent and accurate method SPOT-1D [44] also used predicted contact map information as features and could achieve a significant boost in accuracy. Although these works demonstrate a steady improvement in the published Q8 accuracy over the past few years, the improvements across successive publications are very small. Nevertheless, these small improvements are considered significant given the high complexity of 8-state SS prediction.

Usually the models that focus more on short range dependencies (local context of the amino acid residues) face difficulties in effectively capturing the long range dependencies (interactions between amino acid residues that are close in three-dimensional space, but far from each other in the primary sequence) [22, 27, 45]. Various deep learning based models have been leveraged to handle the long-range interactions by using recurrent or highway networks [28,29], deeper networks with convolutional blocks [30], long short-term memory (LSTM) cells [22,27], whereas the short-range interactions have been handled by convolutional blocks of smaller window size [27, 28, 30]. These methods circumvent some challenging issues in capturing the non-local interactions, but have limitations of their own. Models, using recurrent neural networks to capture long range dependencies, may suffer from *vanishing gradient* or *exploding gradient* problems [46–49]. Moreover, these methods may fail to effectively capture the dependencies when the sequences are very long [50]. Furthermore, as the models grow deeper, the number of parameters also grows which makes it prone to over-fitting. It is also likely that the short range relationships captured in the earlier (shallow) layers may disappear as the models grow deeper [29]. As a result, developing techniques which can capture both long-range and short-range dependencies simultaneously is of utmost importance. Another limiting factor of the deep learning methods is that the high accuracy comes at the expense of high abstraction (less interpretability) due to their black-box nature [51–54]. Although there has been a flurry of recent works towards designing deep learning techniques for bio-molecular data, no notable attempt has been made in developing methods with improved interpretablity and explainability – models that are able to summarize the reasons of the network be-

havior, or produce insights about the causes of their decisions and thus gain trust of users.

In this study, we present SAINT (**S**elf-**A**ttention Augmented **I**nception Inside Inception **N**e**T**work) – a novel method for 8-state SS prediction which uniquely incorporates the *self-attention mechanism* [55] with a state-of-the-art Deep Inception-Inside-Inception (Deep3I) network [30]. We proposed a novel architecture called attention-augmented 3I (2A3I) in order to capture both the local- and long-range interactions. SAINT was compared with a collection of the best alternate methods for Q8 prediction on CASP12 and CASP13 as well as on more recent, challenging and larger test sets (*TEST2016* and *TEST2018*), that were analyzed by a recent and highly accurate method SPOT-1D [44]. SAINT obtained superior Q8 accuracy compared to state-of-the-art predictors on the benchmark datasets – 77.73% accuracy on TEST2016, 76.09% on TEST2018, 74.78% on CASP13, 74.17% in CASP12, and 72.25% on the CASP Free Modeling (FM) targets. SAINT also obtained high precision, recall and $F1$-score for individual states. Moreover, SAINT provides interesting insights regarding the interactions and roles of amino acid residues while forming secondary structures, which help to interpret how the predictions are made. Thus, we have made the following significant contributions: 1) we, for the first time, successfully translated the success of self-attention mechanism from natural language processing to the domain of protein structure prediction, and demonstrated that self-attention improves the accuracy SS prediction, 2) introduced a method which can capture both the short- and long-range dependencies, and offers the best known Q8 accuracy, and 3) improved the interpretability of the black-box deep neural network based methods which are often criticized for lack of interpretability.

2 Approach

2.1 Feature Representation

SAINT takes a protein sequence feature vector $X = (x_1, x_2, x_3, \ldots, x_N)$ as input, where x_i is the vector corresponding to the i^{th} residue, and it returns the protein structure label sequence vector $Y = (y_1, y_2, y_3, \ldots, y_N)$ as output, where y_i is the structure label (one of the eight possible states) of the i^{th} residue. Similar to SPOT-1D-base and MUFOLD-SS, our base model contains 57 features from PSSM profiles, HHM profiles and physicochemical properties. To generate PSSM, PSI-BLAST [56] was run against Uniref90 database [57] with inclusion threshold 0.001 and three iterations. The HHM profiles were generated using HHblits [58] using default parameters against uniprot20_2013_03 sequence database, which can be downloaded from `http://wwwuser.gwdg.de/~compbiol/data/hhsuite/databases/hhsuite_dbs/`. HHblits also generates 7 transition probabilities and 3 local alignment diversity values which we used as features as well. Seven physicochemical properties of each amino acid (e.g., steric parameters (graph-shape index), polarizability, normalized van der Waals volume, hydrophobicity, isoelectric point, helix probability, and sheet probability) were obtained from Meiler *et al.* [59]. So, in our base model, the dimension of x_i is 57 as this is the concatenation of $x\_hhm_i \in \mathbb{R}^{d_{hhm}}$ ($d_{hhm} = 30$), $x\_pssm_i \in \mathbb{R}^{d_{pssm}}$ ($d_{pssm} = 20$), and $x\_physical_i \in \mathbb{R}^{d_{physical}}$($d_{physical} = 7$). Additional features were generated by windowing the predicted contact information as was done in SPOT-1D. The contact maps were generated using SPOT-contact [60] locally and were

used as our features by varying window lengths(the number of preceding or succeeding residues whose pairwise contact information were extracted for a target residue). Our ensemble model constitutes of four different models, that we trained with varying input features: one without the contact maps (base model) and three with different window lengths (10,20, and 50) of the contact-map-based features. The features were normalized to ensure 0 mean and standard deviation of 1 in the training data, similar to SPOT-1D.

2.2 Architecture of SAINT

The architecture of SAINT can be split into three separate discussions: 1) the architecture of our proposed self-attention module, 2) the architecture of the existing inception module and the proposed attention augmented inception module, and finally 3) the overall pipeline of SAINT.

2.2.1 Self-attention module

Attention mechanism implies paying attention to specific parts of input data or features while generating output sequence [55,61]. It calculates a probability distribution over the elements in the input sequence and then takes the weighted sum of those elements based on this probability distribution while generating outputs.

In self-attention mechanism [55, 62, 63], each vector in the input sequence is transformed into three vectors- *query*, *key* and *value*, by three different functions. Each of the output vectors is a weighted sum of the *value* vectors, where the weights are calculated based on the compatibility of the *query* vectors with the *key* vectors by a special function, called *compatibility function* (discussed later in this section).

The self-attention module we designed and augmented with the Deep3I network [30] is inspired from the self-attention module proposed by Vaswani *et al.* [55] and is depicted in Fig. 1. Our self-attention module takes two inputs: 1) the features from the previous inception module or layer, $x \in \mathbb{R}^{d_{protein} \times d_{feature}}$, and 2) position identifiers, $pos\_id \in \mathbb{R}^{d_{protein}}$, where $d_{protein}$ is the length of the protein sequence, and $d_{feature}$ is the length of the feature vector.

Positional Encoding Sub-module. The objective of positional encodings is to inject some information about the relative or absolute positions of the residues in a protein sequence. The *Positional Encoding* $PosEnc_p$ for a position p can be defined as follows [55].

$$PosEnc_{(p,2i)} = sin(p/10000^{2i/d_{feature}}) \tag{1}$$

$$PosEnc_{(p,2i+1)} = cos(p/10000^{2i/d_{feature}}) \tag{2}$$

where i is the dimension. We used such function as it may allow the model to easily learn to attend by relative positions since for any fixed offset k, $PosEnc_{p+k}$ can be represented as a linear function of $PosEnc_p$ [55]. For every position p, $PosEnc_p$ has the dimension $d_{protein} \times d_{feature}$. The output of positional encoding is added with the inputs x, resulting in new representations h (see Eqn. 3) which contain not only the information extracted by the former layers or modules, but also the information about individual positions.

$$h_{pos} = x_{pos} + PosEnc_{pos}. \tag{3}$$

Scaled dot-product attention sub-module. The input features in this sub-module, $h \in \mathbb{R}^{d_{protein} \times d_{feature}}$ are first transformed into three feature spaces Q, K and V, representing *query*, *key* and *value* respectively, in order to compute the scaled dot-product attention, where $Q(h) = W_Q h$, $K(h) = W_K h$, $V(h) = W_V h$. Here W_Q, W_K, W_V are parameter matrices to be learned. Figure 2 shows a schematic diagram of this module.

Among various compatibility functions (e.g. scaled dot-product attention [55], additive attention [61], similarity-attention [64], multiplicative-attention [65], biased general attention [66], etc.), we have chosen the scaled dot-product attention as it showed much promise in case of sequential data. Vaswani *et al.* [55] showed that in practice, the dot-product attention is much faster and space-efficient as it can be implemented using highly optimized matrix multiplication code, though theoretically both dot-product and additive attention have similar complexity. Scaled dot-product $s_{i,j}$ of two vectors h_i and h_j is calculated as shown in Equation 4.

$$s_{i,j} = \frac{Q(h_i)K(h_j)^T}{\sqrt{d_K}} \tag{4}$$

where d_K is the dimension of the feature space K. The numerator of the equation, $Q(h_i)K(h_j)^T$ is the dot product between these two vectors, resulting in the similarity between them in a specific vector space. Here $\sqrt{d_K}$ is the scaling factor which ensures that the result of the dot product does not get prohibitively large for very long sequences.

The attention weights $e \in \mathbb{R}^{d_{protein} \times d_{feature}}$ are calculated as shown in Equation 5, where $e_{j,i}$ represents how much attention have been given to the vector at position i while synthesizing the vector at position j.

$$e_{j,i} = \frac{exp(s_{i,j})}{\sum_{n=1}^{d_{protein}} exp(s_{i,j})} \tag{5}$$

The attention distribution e is multiplied with the feature vectors $V(h)$ and then in order to reduce the internal covariate shift this multiplicand is normalized using *batch normalization* [67], producing g, the output of the scaled dot-product attention sub-module, following the Equation 6.

$$g_j = BatchNorm(\sum_{n=1}^{d_{protein}} e_{j,i} V(h_i)) \tag{6}$$

Here, *BatchNorm(.)* is the batch-normalization function and g_j is the j-th vector in the output sequence of this sub-module. Finally, according to the Equation 7, g is multiplied by a scalar parameter α, the original input feature map x is multiplied by $(1 - \alpha)$ and these two multiplicands are summed to synthesize the final output y.

$$y_i = (\alpha) g_i + (1 - \alpha) x_i \tag{7}$$

where y_i is the ith output and α is a learnable scalar. By introducing weighed sum of g_i and x_i, we give our model the freedom to chose how much weight should be given to each of the features maps, g_i and x_i while generating the output y_i. The optimal value of the parameter α is learnt through back propagation along with the rest of the model.

2.2.2 Attention augmented inception-inside-inception (2A3I) module

A novel deep convolutional neural network architecture, *Inception*, was first introduced by Szegedy *et al.* [68], which demonstrated state-of-the-art performance for image classification and detection. An inception module has several branches, each having one or more convolutional layers. Fang *et al.* used an assembly of inception modules, which they call Inception-inside-Inception (3I module), in their proposed method MUFOLD-SS to predict protein secondary structure. They tried to leverage the inception blocks to retrieve both short-range and long-range dependencies and achieved the best known accuracy at the time. However, convolutional layers cannot capture enough information about long-range similarities or dependencies among feature vectors of a sequence, synthesized by a certain level of the network [69]. In protein secondary structure prediction, this issue leaves more impact on the overall accuracy when the sequence grows in length. Though these types of neural networks that use only convolutional layers need to be deeper to capture the long range dependency, it is often not feasible to add arbitrarily large numbers of layers. Moreover, the authors of MUFOLD-SS showed that using more than two inception-inside-inception modules sequentially does not result into significant increase in the overall accuracy, rather increases the computational expense. Earlier works [19, 20, 22, 27, 70, 71] used Recurrent Neural Network(RNN) based architectures for capturing global features, but incorporating RNN or its derivatives (Gated Recurrent Units (GRU) [72], Long Short Term Memory (LSTM) [73]) inside 3I module would escalate the complexity and computational cost of the model. Therefore, we incorporated the self-attention mechanism to effectively capture both the short-range and long-range dependencies and to bring a better balance between the ability to model long-range dependencies and the computational efficiency. We placed our self attention modules in each branch of the 3I module as shown in Fig. 3. We call this an attention augmented inception-inside-inception (2A3I) module.

2.2.3 Overview of SAINT

A schematic diagram of the overall architecture of SAINT is depicted in Fig. 4. SAINT starts with two consecutive 2A3I modules followed by a self-attention module to supplement the non-local interactions captured by the initial two 2A3I modules. We also observed that this attention module helps achieve faster learning rate. MUFOLD-SS used one convolutional layer with window size 11 after two 3I modules. The level of long-range interactions being captured varies with varying lengths of the window. However, we observed that using window size larger than 11 increases the computational cost without significantly increasing the performance. As a result, we used similar convolutional layer as MUFOLD-SS. However, we included another self-attention module after the convolutional layer to help capture the relations among vectors that the convolutional layer failed to retrieve. The last two dense layers in the MUFOLD-SS were also used in SAINT. However, we placed an attention module in between the two dense layers. We did so to understand how the residues align and interact with each other just before generating the output. This paves the way to have an interpretable deep learning model (as we will discuss in Sec. 3.3.1).

3 Results and Discussion

We performed an extensive evaluation study, comparing SAINT with the state-of-the-art Q8 prediction methods on a collection of publicly available benchmark datasets.

3.1 Dataset

To make a fair comparison with the most recent state-of-the-art methods, especially with SPOT-1D (the most accurate method to date), we used the same training and validation sets that were used in SPOT-1D. However, apart from comparing our model against the most recent and large test sets *TEST2016* and *TEST2018* generated and analyzed by Hanson *et al.* [44, 60], we evaluated SAINT on CASP12, CASP13 and the template free modelling targets of four CASP datasets (CASP10 $\sim$ CASP13). CB513 [41], which is a relatively old, yet widely used benchmark dataset has been excluded from our evaluation as there are many sequences in CB513 with $> 25\%$ sequence similarity to the training set.

The training set contains 10,029 proteins from CullPDB with resolution < 2.5 Å, R-factor < 1.0 and a sequence identity cutoff of 25% according to BlastClust (Altschul *et al.* [56]). The validation set contains 983 proteins from CullPDB with the same specifications applied to training set (see [44] and [60] for more details on these dataset). We provide brief descriptions of the test sets in subsequent sections.

3.1.1 TEST2016

TEST2016 dataset contains 1,213 proteins that were all deposited on PDB between June 2015 and February 2017 with similar parameter settings as the training set and do not contain more than 700 residues. It has less than 25% sequence similarity with the training and validation sets according to BlastClust [56]. It was compiled by Hanson *et al.* [60] and is available at `https://servers.sparks-lab.org/downloads/SPOT-1D-dataset.tar.gz`.

3.1.2 TEST2018

TEST2018 dataset contains 250 high-quality, non-redundant proteins that were all deposited on PDB between January 2018 and July 2018. The dataset was also filtered to remove redundancy at a 25% sequence identity cutoff and to remove proteins having more than 700 residues. It was generated by Hanson *et al.* [60] and is available at `https://servers.sparks-lab.org/downloads/SPOT-1D-dataset.tar.gz`.

3.1.3 CASP

CASP stands for Critical Assessment of protein Structure Prediction. This is an biennial competition for protein structure prediction and a community wide effort to advance the state-of-the-art in modelling protein structure from its amino acid sequences since 1994 [74]. Among the CASP datasets, we took into account the most recent ones CASP13 and CASP12. We removed one domain sequence out of 32 in CASP13 (T0951-D1) and six domain sequences out of 55 in CASP12 as they had more than 25% sequence similarity

to the training set according to CD-HIT [75]. Apart from these, we prepared a dataset comprising the template free modelling (FM) targets in CASP datasets to show the performance of SAINT where the query sequences do not have statistically significant similar protein sequences with known structures. Some of the FM targets had $> 25\%$ sequence similarity with our training set, which were therefore excluded from the test set. Thus, we compiled a test set which we call *CASP-FM* comprising 56 domain sequences: 10 FM targets from CASP13, 22 FM targets from CASP12, 16 FM targets (out of 30) from CASP11, and 8 FM targets (out of 12) from CASP10. The CASP proteins were downloaded from its official website `http://predictioncenter.org/`.

3.2 Method comparison

We compared SAINT with the most recent and accurate Q8 predictors: MUFOLD-SS [30], NetSurfP [43] and SPOT-1D [44]. These state-of-the-art methods have been shown to outperform other popular Q8 predictors, namely, SSPro8 [19], RaptorX-SS8 [40], DeepGSN [26], DeepCNF [25], DCRNN [27], NCCNN [28], CNNH_PSS [29], CBRNN [71], etc.

We evaluated the methods under various evaluation metrics, such as, Q8 accuracy, precision, recall and $F1$-score. We performed Wilcoxon signed-rank test (with $\alpha = 0.05$) to measure the statistical significance of the differences between SAINT and each of the compared state-of-the-art methods.

3.3 Results on benchmark dataset

The comparison of SAINT with the state-of-the-art Q8 structure prediction methods on TEST2016, TEST2018, CASP13, CASP12, and CASPFM is shown in Table 1. To train SAINT and tune necessary hyper-parameters, we have used the same training and validation sets that were used by SPOT-1D. Notably, SPOT-1D is an ensemble of 9 models where each single model uses predicted contact map in addition to other features. SAINT, on the other hand, is an ensemble of only 4 models, 3 of which take advantage of predicted contact map with different windows sizes. Experimental results show that SAINT outperforms all other methods across all the test sets. It is worth mentioning that SAINT's accuracy on the validation set (78.18%) was also better than that of SPOT-1D (77.60%). SPOT-1D's base model, which does not require contact maps as features is also an ensemble of 9 models, whereas SAINT-base model is a single model. Despite being a single model, SAINT-base consistently outperformed SPOT-1D base in TEST2016 and TEST2018. We could not evaluate SPOT-1D base on CASP12, CASP13, and CASPFM as it is not publicly available. From Table 1, it is also evident that SAINT is substantially better than the other recent methods, namely, NetSurfP-2.0 and MUFOLD-SS. Even the base model of SAINT consistently outperformed both NetSurfP-2.0 and MUFOLD-SS. The remarkably large improvement of SAINT over MUFOLD-SS across all the dataset suggests the advantage of augmenting our proposed self-attention mechanism in the Deep3I network used in MUFOLD-SS. Statistical tests (see Table 2) suggests that these improvements of SAINT over other methods are statistically significant ($P << 0.05$).

In addition to the model accuracy, we also investigated the *precision*, *recall* and $F1$-score to obtain better insights on the performances of various methods. Precision, also know as predictivity, denotes the confidence that can be imposed on a prediction. Recall signifies how accurately an algorithm can predict a sample from a particular class. Sometimes an algorithm tends to over-classify which results into high recall but low precision. On the other hand, some algorithms tend to under-classify, preserving the precision at the cost of recall. In order to get an unbiased evaluation of the performance, $F1$-score is considered to be an appropriate measure and has been being used for over 25 years in various domains [76, 77]. Tables 3, 4, and 5 show the precision, recall and $F1$-score on each of the 8 states obtained by SAINT and other methods. These results suggest that SAINT achieves better $F1$-score than other methods on 5 states (out of 8 states), showing that SAINT produced more balanced and meaningful results than other methods. SAINT substantially outperforms other methods on the non-ordinary states [25] such as I, G, S, and T. However, MUFOLD-SS and SPOT-1D achieved slightly better $F1$-score for the 'B' and 'E' states respectively. State 'I' (π-helix) is extremely rare which comprises seven or more residues and is present in 15% of all known protein structures [78]). They are very difficult to predict, but mostly found at functionally important regions such as ligand- and ion-binding sites [78]. Therefore, specialized predictors, such as PiPred [78], is also available that only predicts the π-helix structures. SAINT significantly outperforms SPOT-1D, NetSurfP-2.0, and MUFOLD-SS in predicting π-helix in TEST2016 dataset by correctly predicting 21 out of 47 'I' states and thus achieving a recall of 0.45 for this structure. SAINT's precision for π-helix, on the other hand, is 1. This is remarkable considering the fact that the π-helix specific predictor, PiPred, reports precision and recall of 0.48 and 0.46 respectively on a different dataset which they analyzed in [78].

We analyzed the CASPFM dataset comprising the free modeling targets in the CASP dataset to demonstrate the performance of models on proteins with previously unseen folds. SAINT achieved the best accuracy on CASPFM, suggesting SAINT's superiority in predicting structures of proteins having very low sequence homology with proteins of known structures.

While the advantage of utilizing our proposed self-attention mechanism in the Deep3I framework of MUFOLD-SS is evident from the significant improvement of SAINT over MUFOLD-SS across all the dataset analyzed in this study, we further investigated the efficacy of our proposed attention mechanism in capturing the long-range interactions. We computed the number of non-local interactions per residue for each of the 1,213 proteins in TEST2016, and sorted them in an ascending order. Next, we put them in six equal sized bins $b_1, b_2, \ldots, b_6$ (each containing 202 proteins except for b_6 which contains 203 proteins), where b_1 contains the proteins with the lowest level of non-local interactions and b_6 represents the model condition with the highest level of non-local interactions. We show the Q8 accuracy of SAINT-base and MUFOLD-SS on these model conditions in Table 6 and Fig. 5. Note that, instead of our ensemble model which is more accurate than our base model, we deliberately show the results for our single base model, which uses the same feature set as MUFOLD-SS, and the only difference between them is the self-attention modules introduced in our architecture. These results show that the difference in predictive performance between SAINT-base and MUFOLD-SS significantly increases with increasing levels of non-local interactions. There is no statistically significant difference between them on b_1, but as we increase the level of non-local interactions,

SAINT becomes significantly more accurate than MUFOLD-SS and attains the highest level of improvement on b_6. This clearly indicates that capturing non-local interactions by self-attention is the key factor in the improvement. We also performed the same analyses on other methods (see Fig. 6). The results in Fig. 6 show that the differences among of these methods are not that substantial on the model conditions with low levels of long-range interactions, but the differences become notable as we increase the non-local interactions. SAINT not only achieved the best accuracy, its improvement over other methods increases with increasing amount of long-range interactions as well – suggesting the superiority of our proposed self-attention mechanism compared to CNN+LSTM (used in NetSurfP-2.0) and CNN (ResNet)+BRLSTM (used in SPOT-1D) in terms of capturing the non-local interactions.

In order to demonstrate the efficacy of SAINT and other methods in capturing the continuous structure of a protein, we show the one-dimensional map of the native and predicted secondary structure of a representative protein 5M2PA in TEST2016 (see Fig. 7).

3.3.1 Interpretability

One notable feature of SAINT is that, unlike most of the existing deep learning techniques, it can provide insights on how the architecture is making decisions, especially regarding the long-range interactions. Self-attention alignment matrix has already been used to interpret how different parts of input are dependent on each other while generating the output [55, 79–82], and hence it was used to develop interpretable models [83–86]. Thus, we made an attempt at using attention matrix to capture and provide insight into the long-range interactions. Long-range interactions are crucial for predicting the secondary structure of proteins. For example, a secondary structure state β-strand is stabilized by hydrogen bonds formed with other β-strands that can be far apart from each other in the protein sequence [26].

As mentioned earlier, we placed an attention module just before the last dense layer in the architecture of SAINT. In addition to improving the prediction performance, a motivation behind this attention module has been to introduce some form of interpretability to the deep learning model. Indeed we are able to relate the self-attention alignment score matrix of this attention module to the spatial proximity of a residue with other residues far apart in the primary sequence. In Fig. 8, we show the relation between the spatial proximity and attention scores for a sample protein 5epmD in TEST2016. We selected a short sequence 5epmD (only 33 residues) to easily demonstrate with visualizations how the alignment matrix provides insight about the long-range interactions. We show the distances of the first five residues ('D', 'C', 'L', 'G', 'M') to all other subsequent residues as line graphs and superimpose them on the attention matrix obtained by a single model of SAINT. We choose only the first 5 residues for the sake of readability and clarity of this figure. For this protein, highest attention has been given to the 15-th residue 'K' and 28-th residue 'W', meaning that most of the residues generated the highest attention score with respect to these two residues. Interestingly, these two residues are where the spatial distance lines reach their local minima, indicating a possible turn, bend or contact pair. Being inspired by this, we systematically analyzed the attention matrices and spatial distance graphs of all the 1213 proteins in TEST2016. We consider only those "downslopes" in the spatial proximity line graphs that continue to decrease for at least

three consecutive residues, and thereby ignoring very small decreasing regions which span less than 3 residues. We have observed that, in 93.33% of these proteins, the residues with most attention on them are within a downslope region of another residue's spatial distance line curve. This indicates that the residues with relatively higher attention scores are likely to be spatially closer to some other distant residues in the primary sequence. While these results are promising, especially considering the black-box nature of other deep learning based methods, they should be interpreted with care. The long-range interactions suggested by the attention matrix may contain false positives and false negatives. Higher attention scores do not necessarily guarantee a contact pair, nor is it certain that all the contact pairs will have relatively higher attention scores. More work is required to design an attention mechanism so that the attention matrix is more closely related to the contact map. This is an interesting research avenue which we left as a future work. We believe that this matrix with appropriate modifications will be useful to understand the complex relationship between the primary sequence and various structural and functional properties of proteins.

3.3.2 Running time

SAINT is much faster than the best alternate method SPOT-1D. For generating the structures of 1,213 protein chains in TEST2016, given the necessary input files, SAINT took approximately 360 ± 5 seconds whereas SPOT-1D took approximately $2,485 \pm 5$ seconds on our local machine (Intel corei7-7700 CPU 3.60 GHz (4 cores), 16GB RAM, NVIDIA GeForce GTX 1070 GPU). Under the same settings, SAINT took approximately 197 ± 5 seconds to generate secondary structures for the 250 proteins in TEST2018, whereas SPOT-1D took approximately 668 ± 5 seconds. Since both these methods use the same input files for feature generation, this substantial difference in running time can be attributed to the efficiency of our attention based method over the LSTM network-based model used in SPOT-1D.

4 Conclusions

We have presented SAINT, a highly accurate, fast, and interpretable method for 8-state SS prediction. We demonstrate for the first time that the self-attention mechanism proposed by Vaswani *et al.* [55] is a valuable tool to apply in the structural analyses of proteins. Another earlier type of attention mechanism proposed by Bahdanau *et al.* [61] coupled with recurrent neural network (RNN) based encoder-decoder architectures achieved state-of-the-art performance on various natural language processing tasks (e.g. neural machine translation [65,87], question answering task [88,89], text summarization [90,91], document classification [92,93], sentiment classification [94,95], etc.). As proteins are also sequences similar to sentences in a language, this type of architecture is expected do well in protein secondary structure prediction as well. However, previous attempts [96] on using attention with LSTM based encoder-decoder only achieved 68.4% accuracy on CB513 dataset which is significantly worse than the performance of MUFOLD-SS (70.63% on CB513) [30]. In this study, we have used the self-attention mechanism in a unique way and proposed a novel attention augmented 3I module (2A3I module) and achieved notable success. We have used the self-attention mechanism to retrieve the relation between

vectors that lay far from each other in a sequence. As self-attention mechanism looks at a single vector and measures its similarity or relationship with all other vectors in the same sequence, it does not need to encode all the information in a sequence into a single vector like recurrent neural networks. This reduces the loss of contextual information for long sequences.

SAINT contributes towards simultaneously capturing the local and non-local dependencies among the amino acid residues. Unlike some of the existing deep learning methods, SAINT can capture the long-range dependencies without using computationally expensive recurrent networks or convolution networks with large window sizes. SAINT was assessed for its performance against the state-of-the-art 8-state SS prediction methods on a collection of widely used benchmark dataset. Experimental results suggest that SAINT consistently improved upon the best existing methods across various widely used benchmark dataset.

One of the most significant conclusions from the demonstrated experimental results is that appropriate use of self-attention mechanism can significantly boost the performance of deep neural networks and is capable of producing results which rank SAINT at the very top of the current SS prediction methods. Thus, the idea of applying self-attention mechanism can be applied to predicting various other protein attributes (e.g., torsion angles, turns, etc. [97]) as well. Therefore, we believe SAINT advances the state-of-the-art in this domain, and will be considered as a useful tool for predicting the secondary structures of proteins.

References

[1] Qian Jiang, Xin Jin, Shin-Jye Lee, and Shaowen Yao. Protein secondary structure prediction: A survey of the state of the art. *Journal of Molecular Graphics and Modelling*, 76:379–402, 2017.

[2] Christian B Anfinsen. Principles that govern the folding of protein chains. *Science*, 181(4096):223–230, 1973.

[3] David Baker and Andrej Sali. Protein structure prediction and structural genomics. *Science*, 294(5540):93–96, 2001.

[4] Ken A Dill, S Banu Ozkan, M Scott Shell, and Thomas R Weikl. The protein folding problem. *Annu. Rev. Biophys.*, 37:289–316, 2008.

[5] Philip Bradley, Kira MS Misura, and David Baker. Toward high-resolution de novo structure prediction for small proteins. *Science*, 309(5742):1868–1871, 2005.

[6] Linus Pauling, Robert B Corey, and Herman R Branson. The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences*, 37(4):205–211, 1951.

[7] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637, 1983.

[8] Ning Qian and Terrence J Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202(4):865–884, 1988.

[9] L Howard Holley and Martin Karplus. Protein secondary structure prediction with a neural network. *Proceedings of the National Academy of Sciences*, 86(1):152–156, 1989.

[10] Burkhard Rost and Chris Sander. Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232(2):584–599, 1993.

[11] Markéta J Zvelebil, Geoffrey J Barton, William R Taylor, and Michael JE Sternberg. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *Journal of Molecular Biology*, 195(4):957–961, 1987.

[12] David T Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999.

[13] Hyunsoo Kim and Haesun Park. Protein secondary structure prediction based on an improved support vector machines approach. *Protein Engineering*, 16(8):553–560, 2003.

[14] Jonathan J Ward, Liam J McGuffin, Bernard F. Buxton, and David T. Jones. Secondary structure prediction with support vector machines. *Bioinformatics*, 19(13):1650–1655, 2003.

[15] Jian Guo, Hu Chen, Zhirong Sun, and Yuanlie Lin. A novel method for protein secondary structure prediction using dual-layer svm and profiles. *PROTEINS: Structure, Function, and Bioinformatics*, 54(4):738–743, 2004.

[16] Wei Chu, Zoubin Ghahramani, and David L Wild. A graphical model for protein secondary structure prediction. In *Proceedings of the twenty-first international conference on Machine learning*, page 21. ACM, 2004.

[17] Kiyoshi Asai, Satoru Hayamizu, and Ken'ichi Handa. Prediction of protein secondary structure by the hidden markov model. *Bioinformatics*, 9(2):141–146, 1993.

[18] Zafer Aydin, Yucel Altunbasak, and Mark Borodovsky. Protein secondary structure prediction for a single-sequence using hidden semi-markov models. *BMC Bioinformatics*, 7(1):178, 2006.

[19] Gianluca Pollastri, Darisz Przybylski, Burkhard Rost, and Pierre Baldi. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235, 2002.

[20] Jinmiao Chen and Narendra Chaudhari. Cascaded bidirectional recurrent neural networks for protein secondary structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(4):572–582, 2007.

[21] Claudio Mirabello and Gianluca Pollastri. Porter, paleale 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, 29(16):2056–2058, 2013.

[22] Rhys Heffernan, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, 33(18):2842–2849, 2017.

[23] Matt Spencer, Jesse Eickholt, and Jianlin Cheng. A deep learning network approach to ab initio protein secondary structure prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 12(1):103–112, 2015.

[24] Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal, and Yaoqi Zhou. Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Briefings in Bioinformatics*, 19(3):482–494, 2016.

[25] Sheng Wang, Jian Peng, Jianzhu Ma, and Jinbo Xu. Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports*, 6:18962, 2016.

[26] Jian Zhou and Olga G. Troyanskaya. Deep supervised and convolutional generative stochastic network for protein secondary structure prediction. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages 745–753, 2014.

[27] Zhen Li and Yizhou Yu. Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2560–2567. AAAI Press, 2016.

[28] Akosua Busia and Navdeep Jaitly. Next-step conditioned deep convolutional neural networks improve protein secondary structure prediction. *arXiv preprint arXiv:1702.03865*, 2017.

[29] Jiyun Zhou, Hongpeng Wang, Zhishan Zhao, Ruifeng Xu, and Qin Lu. Cnnh_pss: protein 8-class secondary structure prediction by convolutional neural network with highway. *BMC Bioinformatics*, 19(4):60, 2018.

[30] Chao Fang, Yi Shang, and Dong Xu. Mufold-ss: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 86(5):592–598, 2018.

[31] Sujun Hua and Zhirong Sun. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *Journal of Molecular Biology*, 308(2):397–407, 2001.

[32] Scott C Schmidler, Jun S Liu, and Douglas L Brutlag. Bayesian segmentation of protein secondary structure. *Journal of Computational Biology*, 7(1-2):233–248, 2000.

[33] Laurens Van Der Maaten, Max Welling, and Lawrence Saul. Hidden-unit conditional random fields. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 479–488, 2011.

[34] Pierre Baldi, Søren Brunak, Paolo Frasconi, Giovanni Soda, and Gianluca Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

[35] Christophe N Magnan and Pierre Baldi. Sspro/accpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, 30(18):2592–2597, 2014.

[36] Jie Hou, Zhiye Guo, and Jianlin Cheng. Dnss2: improved ab initio protein secondary structure prediction using advanced deep learning architectures. *bioRxiv*, page 639021, 2019.

[37] GN t Ramachandran and V Sasisekharan. Conformation of polypeptides and proteins. In *Advances in protein chemistry*, volume 23, pages 283–437. Elsevier, 1968.

[38] Ashraf Yaseen and Yaohang Li. Template-based c8-scorpion: a protein 8-state secondary structure prediction method using structural information and context-based features. *BMC bioinformatics*, 15(S8):S3, 2014.

[39] Maya Topf, Matthew L Baker, Marc A Marti-Renom, Wah Chiu, and Andrej Sali. Refinement of protein structures by iterative comparative modeling and cryoem density fitting. *Journal of molecular biology*, 357(5):1655–1668, 2006.

[40] Zhiyong Wang, Feng Zhao, Jian Peng, and Jinbo Xu. Protein 8-class secondary structure prediction using conditional neural fields. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 109–114. IEEE, 2010.

[41] James A Cuff and Geoffrey J Barton. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Structure, Function, and Bioinformatics*, 34(4):508–519, 1999.

[42] Yanbu Guo, Weihua Li, Bingyi Wang, Huiqing Liu, and Dongming Zhou. Deepaclstm: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC Bioinformatics*, 20(1):341, 2019.

[43] Michael Schantz Klausen, Martin Closter Jespersen, Henrik Nielsen, Kamilla Kjaergaard Jensen, Vanessa Isabell Jurtz, Casper Kaae Soenderby, Morten Otto Alexander Sommer, Ole Winther, Morten Nielsen, Bent Petersen, et al. Netsurfp-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins: Structure, Function, and Bioinformatics*, 87(6):520–527, 2019.

[44] Jack Hanson, Kuldip Paliwal, Thomas Litfin, Yuedong Yang, and Yaoqi Zhou. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. *Bioinformatics*, 35(14):2403–2410, 2019.

[45] Jack Hanson, Yuedong Yang, Kuldip Paliwal, and Yaoqi Zhou. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, 33(5):685–692, 2016.

[46] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *CoRR, abs/1211.5063*, 2, 2012.

[47] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[48] Sekitoshi Kanai, Yasuhiro Fujiwara, and Sotetsu Iwamura. Preventing gradient explosions in gated recurrent units. In *Advances in neural information processing systems*, pages 435–444, 2017.

[49] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232, 2016.

[50] Yoshua Bengio, Patrice Simard, Paolo Frasconi, et al. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks and Learning Systems*, 5(2):157–166, 1994.

[51] Quan-shi Zhang and Song-Chun Zhu. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology & Electronic Engineering*, 19(1):27–39, 2018.

[52] Supriyo Chakraborty, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, Alun Preece, Simon Julier, Raghuveer M Rao, et al. Interpretability of deep learning models: a survey of results. In *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation*, pages 1–6. IEEE, 2017.

[53] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

[54] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1885–1894. JMLR. org, 2017.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[56] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402, 1997.

[57] UniProt Consortium. The universal protein resource (uniprot). *Nucleic Acids Research*, 36(suppl_1):D190–D195, 2007.

[58] Michael Remmert, Andreas Biegert, Andreas Hauser, and Johannes Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173, 2012.

[59] Jens Meiler, Michael Müller, Anita Zeidler, and Felix Schmäschke. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Molecular modeling annual*, 7(9):360–369, 2001.

[60] Jack Hanson, Kuldip Paliwal, Thomas Litfin, Yuedong Yang, and Yaoqi Zhou. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, 34(23):4039–4045, 2018.

[61] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[62] Jianpeng Cheng, Li Dong, and Mirella Lapata. Long short-term memory-networks for machine reading. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 551–561, 2016.

[63] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, 2016.

[64] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[65] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[66] Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. Iterative alternating neural attention for machine reading. *arXiv preprint arXiv:1606.02245*, 2016.

[67] Sergey Ioffe and Christian Szegedy. Batch normalization: accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37*, pages 448–456. JMLR. org, 2015.

[68] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[69] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *Stat*, 1050:21, 2018.

[70] Søren Kaae Sønderby and Ole Winther. Protein secondary structure prediction with long short term memory networks. *arXiv preprint arXiv:1412.7828*, 2014.

[71] Yanbu Guo, Bingyi Wang, Weihua Li, and Bei Yang. Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. *Journal of Bioinformatics and Computational Biology*, 16(05):1850021, 2018.

[72] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[73] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[74] Patrice Koehl and Michael Levitt. A brighter future for protein structure prediction. *Nature Structural Biology*, 6(2):108, 1999.

[75] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, 2010.

[76] Yutaka Sasaki et al. The truth of the f-measure. *Teach Tutor Mater*, 1(5):1–5, 2007.

[77] Erik F Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.

[78] Jan Ludwiczak, Aleksander Winski, Antonio Marinho da Silva Neto, Krzysztof Szczepaniak, Vikram Alva, and Stanislaw Dunin-Horkawicz. Pipred–a deep-learning method for prediction of π-helices in protein sequences. *Scientific Reports*, 9(1):6888, 2019.

[79] Reza Ghaeini, Xiaoli Fern, and Prasad Tadepalli. Interpreting recurrent and attention-based neural models: a case study on natural language inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4952–4957, 2018.

[80] Jaesong Lee, Joong-Hwi Shin, and Jun-Seok Kim. Interactive visualization and manipulation of attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 121–126, 2017.

[81] Tamer Alkhouli and Hermann Ney. Biasing attention-based recurrent neural networks using external alignment information. In *Proceedings of the Second Conference on Machine Translation*, pages 108–117, 2017.

[82] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[83] Mostafa Karimi, Di Wu, Zhangyang Wang, and Yang Shen. Deepaffinity: interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks. *Bioinformatics*, 35(18):3329–3338, 2019.

[84] Chen Chen, Jie Hou, Xiaowen Shi, Hua Yang, James A Birchler, and Jianlin Cheng. Interpretable attention model in transcription factor binding site prediction with deep neural networks. *bioRxiv*, page 648691, 2019.

[85] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017.

[86] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, pages 3504–3512, 2016.

[87] Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, 2016.

[88] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*, pages 2397–2406, 2016.

[89] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297, 2016.

[90] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, 2015.

[91] Abigail See, Peter J Liu, and Christopher D Manning. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, 2017.

[92] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[93] Nikolaos Pappas and Andrei Popescu-Belis. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025, 2017.

[94] Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, 2016.

[95] Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615, 2016.

[96] Iddo Drori, Isht Dwivedi, Pranav Shrestha, Jeffrey Wan, Yueqi Wang, Yunchu He, Anthony Mazza, Hugh Krogh-Freeman, Dimitri Leggas, Kendal Sandridge, et al. High quality prediction of protein q8 secondary structure by diverse neural network architectures. *arXiv preprint arXiv:1811.07143*, 2018.

[97] Chao Fang, Zhaoyu Li, Dong Xu, and Yi Shang. Mufold-ssw: A new web server for predicting protein secondary structures, torsion angles, and turns. *Bioinformatics*, 2019.

[98] Warren L DeLano et al. Pymol: An open-source molecular graphics tool. *CCP4 Newsletter on Protein Crystallography*, 40(1):82–92, 2002.

Tables

Method	TEST2016	TEST2018	CASP13	CASP12	CASPFM
SAINT[1]	**77.73**	**76.10**	**74.78**	**74.17**	**72.25**
SPOT-1D[1]	77.10	75.41	73.91	73.67	71.85
SAINT-base	76.23	74.48	73.5	71.78	70.00
SPOT-1D-base[1,2]	76.03	74.26	-	-	-
NetSurfP-2.0	75.68	73.04	72.85	71.43	70.16
MUFOLD-SS	75.56	73.66	73.44	71.44	70.21

[1] Indicates ensemble model
[2] Not publicly available. Results reported by SPOT-1D [44].

Table 1: A comparison of the Q8 accuracy (%) obtained by SAINT and other state-of-the-art methods on TEST2016 and TEST2018 dataset. Best results for each benchmark dataset are shown in bold.

Method	TEST2016 (1213)	TEST2018 (250)	CASP13 (31)	CASP12 (49)	CASPFM (56)
SPOT-1D	8.168e-27	3.893e-5	0.101	0.0345	0.0791
NetSurfP-2.0	2.607e-57	3.258e-18	0.179	1.55e-6	0.0001
MUFOLD-SS	1.531e-88	3.145e-21	0.179	6.51e-5	0.005

Table 2: **Statistical significance of the Q8 accuracy between SAINT and other state-of-the-art methods.** The numbers of protein chains or domains in these datasets are shown in parentheses. We show the p-values using a Wilcoxon signed rank test.

Q8 Label	SAINT	SPOT-1D	NetSurfP-2.0	MUFOLD-SS
H	0.879	0.884	**0.885**	0.868
B	**0.76**	0.671	0.65	0.609
E	0.843	**0.852**	0.822	0.85
G	**0.581**	0.547	0.536	0.519
I	**1**	1	0.044	0.857
T	**0.663**	0.641	0.615	0.631
S	**0.639**	0.624	0.579	0.589
C	**0.648**	0.631	0.613	0.607

Table 3: Predictive precision on each of the 8 states obtained by SAINT and other state-of-the-art methods on TEST2016 dataset.

Q8 Label	SAINT	SPOT-1D	NetSurfP-2.0	MUFOLD-SS
H	**0.948**	0.941	0.933	0.943
B	0.104	0.097	0.07	**0.115**
E	0.887	0.878	**0.903**	0.842
G	**0.39**	0.375	0.334	0.348
I	**0.447**	0.128	0.426	0.383
T	**0.618**	0.612	0.585	0.586
S	**0.367**	0.337	0.278	0.313
C	0.731	**0.741**	0.704	0.727

Table 4: Recall on each of the 8 states obtained by SAINT and other state-of-the-art methods on TEST2016 dataset.

Q8 Label	Frequency	SAINT	SPOT-1D	NetSurfP-2.0	MUFOLD-SS
H	98139	**0.912**	0.911	0.908	0.904
B	3018	0.183	0.169	0.126	**0.193**
E	62657	0.864	**0.865**	0.861	0.846
G	10770	**0.467**	0.445	0.412	0.417
I	47	**0.618**	0.227	0.079	0.529
T	32297	**0.639**	0.626	0.599	0.608
S	23466	**0.466**	0.438	0.376	0.409
C	57483	**0.687**	0.682	0.655	0.662

Table 5: $F1$-score on each of the 8 states obtained by SAINT and other state-of-the-art methods on TEST2016 dataset.

Non-local contacts per residue	**Q8 Accuracy (%) of SAINT-base**	**Q8 Accuracy (%) of MUFOLD-SS**	**Accuracy Difference**	***p*-value**
0-0.61	83.01	83.05	-0.04	0.721287
0.61-0.99	78.77	78.46	0.31	0.01036
0.99-1.24	75.80	75.37	0.43	0.004
1.24-1.45	75.79	75.14	0.65	0.0003
1.45-1.64	75.46	74.60	0.86	0.0001
1.64-2.70	73.79	72.63	1.16	1.36e-6

Table 6: **Accuracy of SAINT(base) and MUFOLD-SS under various levels of non-local interactions.** 1,213 proteins in TEST2016 were divided into 6 disjoint bins each having 202 proteins except the last one which had 203 proteins. The binning was based on the number of non-local contacts per residue in the proteins.

Figures

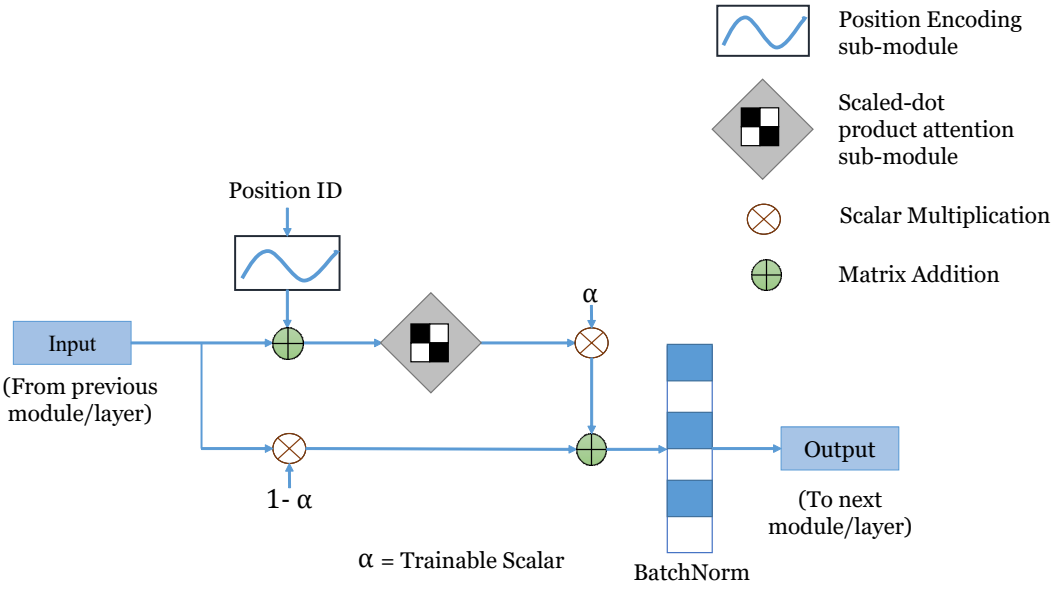


Figure 1: Architecture of the self-attention module used in SAINT.

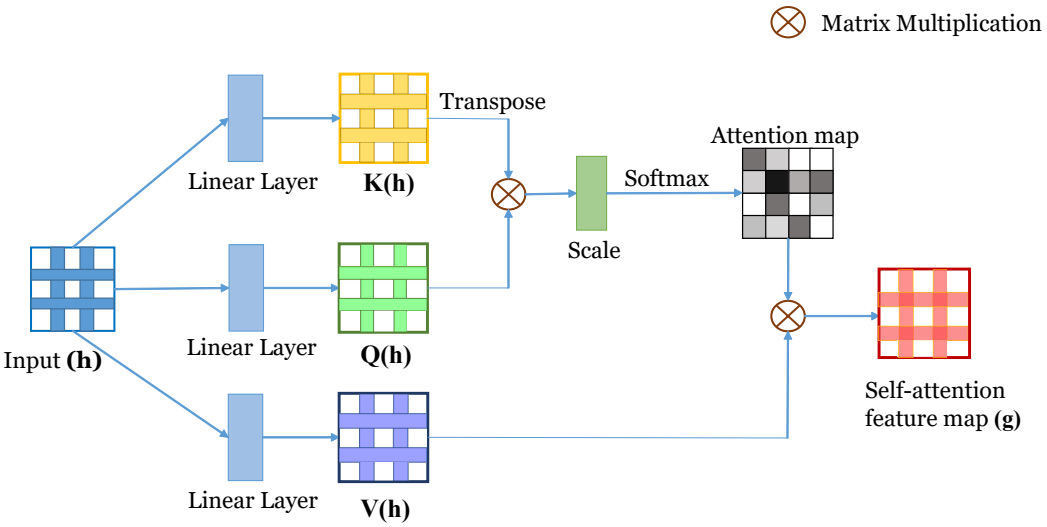


Figure 2: Architecture of the scaled dot-product attention sub-module.

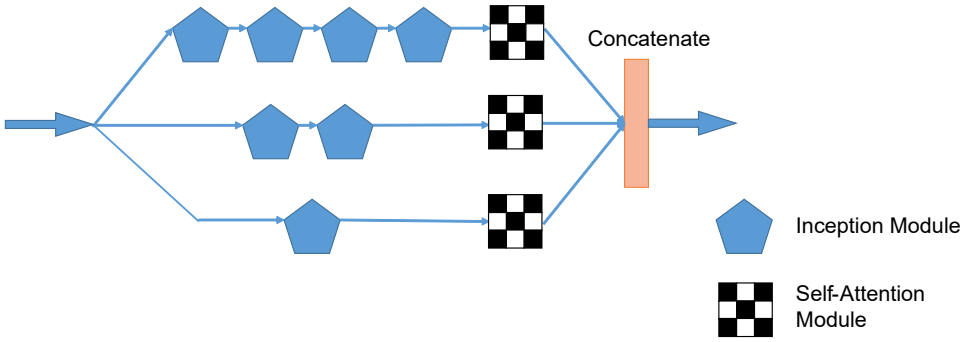


Figure 3: Architecture of our proposed 2A3I module by augmenting self-attention within the inception-inside-inception (3I) network.

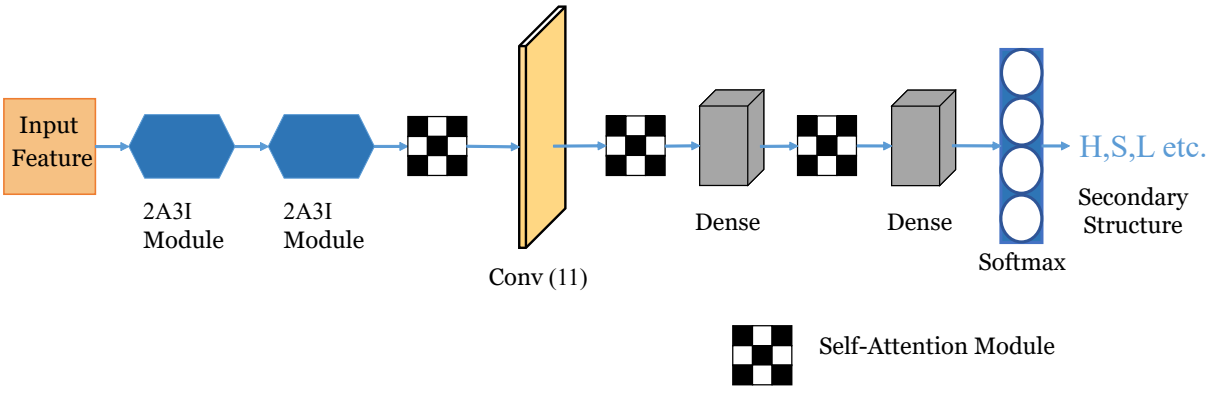


Figure 4: **A schematic diagram of the overall architecture of SAINT.** It comprises two 2A3I modules, three self-attention modules, convolutional layers with window size 11 and two dense layers.

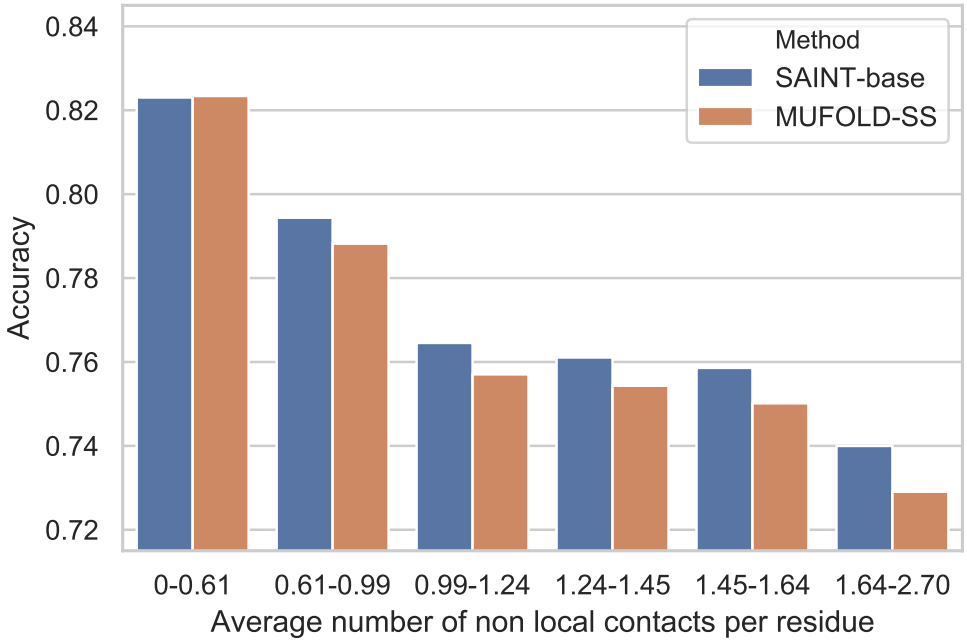


Figure 5: **Accuracy of SAINT-base and MUFOLD-SS under various levels of non-local interactions**. We show the results on the TEST2016 test set using six bins of proteins as shown in Table 6.

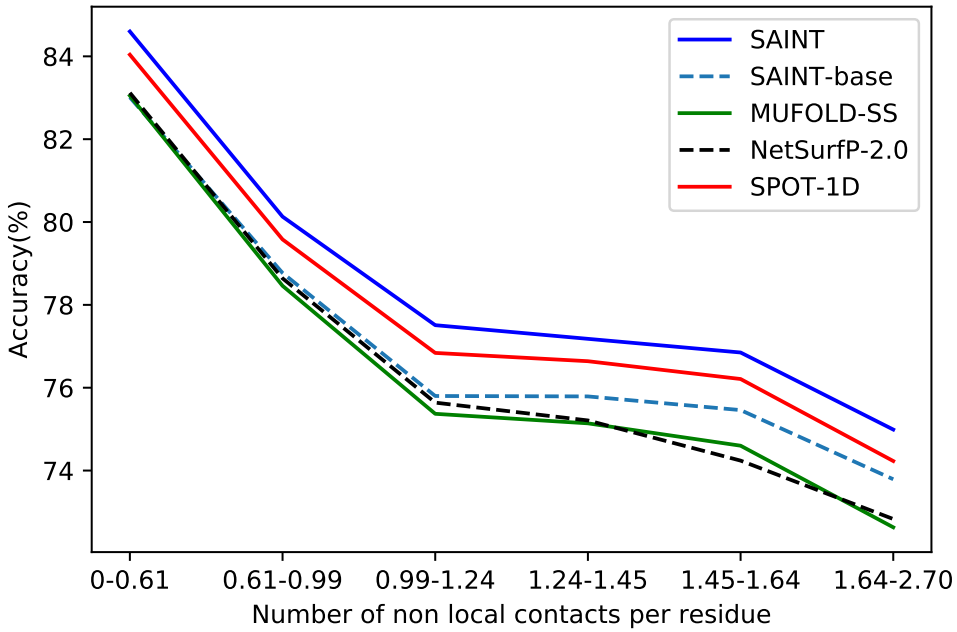
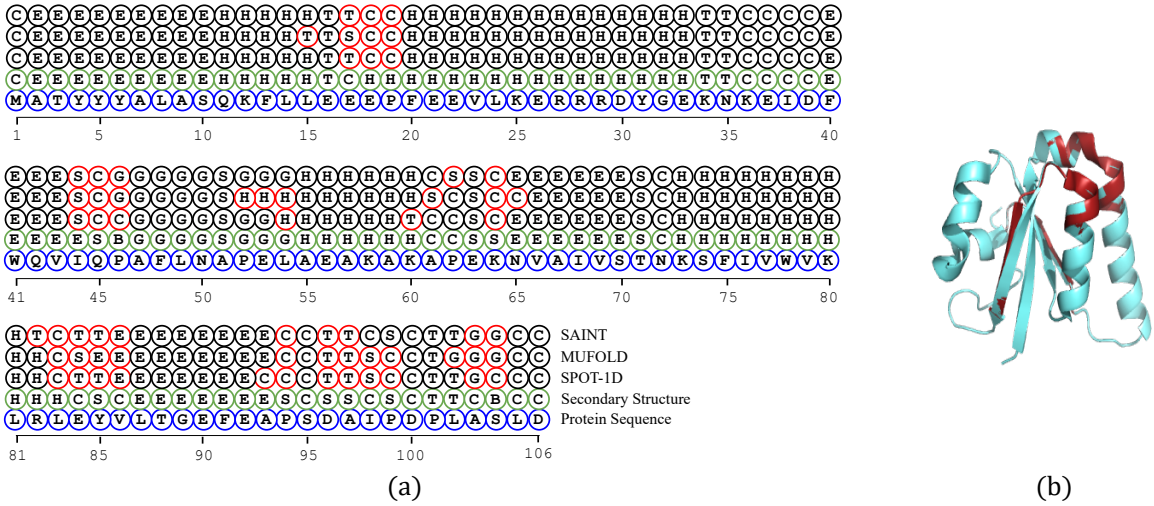


Figure 6: **Accuracy of SAINT, SPOT-1D, NetSurfP-2.0 and MUFOLD-SS as a function of the average number of non-local interactions per residue.** We show the results on the six bins as shown in Table 6.



(a) (b)

Figure 7: **Structure prediction on 5M2PA protein chain by various methods.** (a) One-dimensional map of the native structure of 5M2PA and the predicted structures by various methods. (b) Superposition of the structures predicted by SAINT (cyan) with the structures obtained from PDB (red) for 5M2PA. This image is generated in Pymol [98]. Since Pymol does not differentiate between all the 8 distinct states, we translated the 8-state structure to 3-state structure according to the Rost and Sander scheme [10].

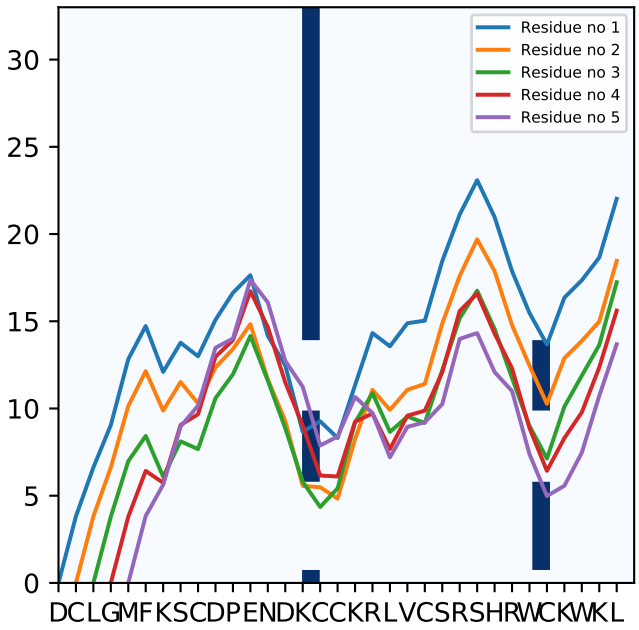


Figure 8: **Demonstration of the interpretability of SAINT using the attention map.** Spatial distances of the first five residues ('D', 'C', 'L', 'G', 'M') in 5epmD to all other subsequent residues are shown by line graphs and they are superimposed on the attention matrix. Deeper hue on the 15-th residue 'K' and the 28-th residue 'W' indicates higher attention scores.